

## Claims

### What is claimed is:

- 1           1.    A method of determining the language of a textual  
2                    passage, the method comprising the steps of:
  - 3                    (a)   parsing said textual passage into a plurality of  
4                            n-grams;
  - 5                    (b)   comparing each of said n-grams with a plurality  
6                            of databases, wherein each of said databases  
7                            comprises a list of n-grams associated with a  
8                            specific language;
  - 9                    (c)   determining an initial weight for each of said n-  
10                           grams, per language, by calculating the frequency  
11                           with which each of said n-grams appears in each  
12                           of said databases and dividing said frequency by  
13                           the total number of n-grams in said respective  
14                           database;
  - 15                   (d)   determining the number of said databases within  
16                           which each of said n-grams appear;
  - 17                   (e)   altering said initial weight for each of said n-  
18                           grams by multiplying said initial weight with the  
19                           inverse of said number of databases within which  
20                           each of said n-grams appear;
  - 21                   (f)   producing the weight of each language over the  
22                           text passage by calculating, per language, the  
23                           sum over each n-gram in the text passage of the  
24                           products of the number of times that that n-gram  
25                           appears in the text passage and the language-

26                   specific altered weight calculated in step (e)  
27                   for that n-gram;

28           (g)    sorting the list of per language passage weights  
29                   from step (f) in decreasing order, returning the  
30                   most likely language for the text passage as the  
31                   first element (highest weight) in the list.

1           2.    The method of claim 1 wherein the step of determining  
2                   an initial weight for each of said n-grams, per  
3                   language, comprises the steps of calculating the  
4                   frequency with which each of said n-grams appears in  
5                   each of said databases and dividing said frequency by  
6                   the total number of n-grams in said respective  
7                   database.

1           3.    The method of claim 1 wherein said n-grams have a size  
2                   selected from the group consisting of bi-grams, tri-  
3                   grams, and quad-grams.

1           4.    The method of claim 1 wherein said n-grams are  
2                   anchored n-grams.

1           5.    The method of claim 1 wherein said n-grams are  
2                   replacement-type n-grams.

1           6.    The method of claim 1 wherein said n-grams are any  
2                   combination of n-grams, including anchored n-grams

3 and/or replacement-type n-grams, and/or n-grams of  
4 different lengths.

5  
1 7. The method of claim 1 wherein said textual passage  
2 comprises 20 or more words.

1 8. The method of claim 1 wherein said textual passage  
2 comprises 40 or more words.

1 9. method of determining the language of a textual  
2 passage, the method comprising the steps of:

3 (a) filtering a plurality of short words from a  
4 textual passage;

5 (b) comparing each of said short words against a  
6 plurality of databases, wherein each of said  
7 databases comprises a list of short words  
8 associated with a different language;

9 (c) determining an initial weight for each of said  
10 short words, per language, by calculating the  
11 frequency with which each of said short words  
12 appears in each of said databases and dividing  
13 said frequency by the total number of short words  
14 in said respective database;

15 (d) determining the number of said databases within  
16 which each of said short words appear;

17 (e) altering said initial weight for each of said  
18 short words by multiplying said initial weight

19                   with the inverse of said number of databases  
20                   within which each of said short words appear;  
  
21           producing the weight of each language over the text  
22           passage by calculating, per language, the sum  
23           over each short word in the text passage of the  
24           products of the number of times that that short  
25           word appears in the text passage and the  
26           language-specific altered weight calculated in  
27           step (e) for that short word;  
  
28           (g)    sorting the list of per language passage weights  
29           from step (f) in decreasing order, returning the  
30           most likely language for the text passage as the  
31           first element (highest weight)   in the list.

1       10.   A method of determining the language of a textual  
2       passage, the method comprising the steps of:

3           (a)    filtering a plurality of short words from a  
4           textual passage and parsing said textual passage  
5           into a plurality of n-grams;  
  
6           (b)    comparing each of said n-grams and said short  
7           words against a plurality of databases, wherein  
8           each of said databases comprises a list of n-  
9           grams and short words associated with a different  
10          language;  
  
11          (c)    determining an initial weight for each of said n-  
12          grams and said short words, per language;

13 (d) determining the number of said databases within  
14 which each of said n-grams and said short words  
15 appear;

16 (e) altering said initial weight for each of said n-  
17 grams and said short words by multiplying said  
18 initial weight with the inverse of said number of  
19 databases within which each of said n-grams and  
20 said short words appear;

21 producing the weight of each language over the text  
22 passage by calculating, per language, the sum  
23 over each short word and each n-gram in the text  
24 passage of the products of the number of times  
25 that that short word or n-gram appears in the  
26 text passage and the language-specific altered  
27 weight calculated in step (e) for that short word  
28 or n-gram;

29 (g) sorting the list of per language passage weights  
30 from step (f) in decreasing order, returning the  
31 most likely language for the text passage as the  
32 first element (highest weight) in the list.

1 11. A system for determining the language of a textual  
2 passage, comprising:

3 a central processing unit coupled to a memory system  
4 and a display , wherein said central processing unit  
5 operates according to a program retrieved from said  
6 memory system, wherein said program includes the steps  
7 of;

8 (a) receiving a textual passage;

- 9 (b) parsing said textual passage into a plurality of  
10 n-grams;
- 11 (c) comparing each of said n-grams against a  
12 plurality of databases, wherein each of said  
13 databases comprises a list of n-grams associated  
14 with a different language;
- 15 (d) assigning an initial weight to each of said n-  
16 grams, per language, by calculating the frequency  
17 with which each of said n-grams appears in each  
18 of said databases and dividing said frequency by  
19 the total number of n-grams in said respective  
20 database;
- 21 (e) calculating the number of said databases within  
22 which each of said n-grams appear;
- 23 (f) altering said initial weight assigned to each of  
24 said n-grams by multiplying said initial weight  
25 with the inverse of said number of databases  
26 within which each of said n-grams appear;
- 27 (g) producing the weight of each language over the  
28 text passage by calculating, per language, the  
29 sum over each n-gram in the text passage of the  
30 products of the number of times that that n-gram  
31 appears in the text passage and the language-  
32 specific altered weight calculated in step (f)  
33 for that n-gram;
- 34 (h) sorting the list of per language passage weights  
35 from step (g) in decreasing order, returning the

36                   most likely language for the text passage as the  
37                   first element (highest weight) in the list.

1       12.   The system of claim 11 further comprising a scanner  
2            and an optical character recognition device, wherein  
3            said scanner and said optical character recognition  
4            device are connected to said central processing unit,  
5            wherein said program receives a textual passage from a  
6            document scanned by said scanner.

1       13.   The system of claim 11 wherein said program comprises  
2            a user interface that allows a user to enter said  
3            textual passage.

1       14.   The system of claim 13 wherein said user interface is  
2            a graphical user interface.

1       15.   The system of claim 13 wherein said user interface  
2            displays the identified language.

1       16.   The system of claim 11 wherein said program comprises  
2            a user interface that allows a user to enter a Uniform  
3            Resource Locator that contains said textual passage.